REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
	New Reprint	-

4. TITLE AND SUBTITLE
Exploiting Sparsity in Hyperspectral ImageClassification via Graphical Models

5a. CONTRACT NUMBER
W911NF-11-1-0245

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Umamahesh Srinivas, Yi Chen, Vishal Monga, Nasser M. Nasrabadi, Trac D. Tran

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
Johns Hopkins University
3400 N Charles Street
Maryland Hall
Batlimore, MD 21218 -2608

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
60291-MA.11

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
A significant recent advance in hyperspectral image (HSI) classification relies on the observation that the spectral signature of a pixel can be represented by a sparse linear combination of training spectra from an overcomplete dictionary. A spatiospectral notion of sparsity is further captured by developing

15. SUBJECT TERMS
Classification, hyperspectral imagery, joint sparsitymodel, probabilistic graphical models, sparse representation,spatial correlation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Trac Tran
UU	UU	UU	UU		19b. TELEPHONE NUMBER 410-516-7416

Report Title

Exploiting Sparsity in Hyperspectral ImageClassification via Graphical Models

ABSTRACT

A significant recent advance in hyperspectral image (HSI) classification relies on the observation that the spectral signature of a pixel can be represented by a sparse linear combination of training spectra from an overcomplete dictionary. A spatiospectral notion of sparsity is further captured by developing a joint sparsity model, wherein spectral signatures of pixels in a local spatial neighborhood (of the pixel of interest) are constrained to be represented by a common collection of training spectra, albeit with different weights. A challenging open problem is to effectively capture the class conditional correlations between these multiple sparse representations corresponding to different pixels in the spatial neighborhood. We propose a probabilistic graphical model framework to explicitly mine the conditional dependences between these distinct sparse features. Our graphical models are synthesized using simple tree structures which can be discriminatively learnt (even with limited training samples) for classification. Experiments on benchmark HSI data sets reveal significant improvements over existing approaches in classification rates as well as robustness to choice of training.

REPORT DOCUMENTATION PAGE (SF298)
(Continuation Sheet)

Continuation for Block 13

ARO Report Number 60291.11-MA
Exploiting Sparsity in Hyperspectral ImageClass ...

Block 13: Supplementary Note

Approved for public release; distribution is unlimited.

Exploiting Sparsity in Hyperspectral Image Classification via Graphical Models

Umamahesh Srinivas, *Student Member, IEEE*, Yi Chen, *Student Member, IEEE*, Vishal Monga, *Senior Member, IEEE*, Nasser M. Nasrabadi, *Fellow, IEEE*, and Trac D. Tran, *Senior Member, IEEE*

***Abstract*—A significant recent advance in hyperspectral image (HSI) classification relies on the observation that the spectral signature of a pixel can be represented by a *sparse* linear combination of training spectra from an overcomplete dictionary. A spatiospectral notion of sparsity is further captured by developing a joint sparsity model, wherein spectral signatures of pixels in a local spatial neighborhood (of the pixel of interest) are constrained to be represented by a common collection of training spectra, albeit with different weights. A challenging open problem is to effectively capture the *class conditional correlations* between these multiple sparse representations corresponding to different pixels in the spatial neighborhood. We propose a probabilistic graphical model framework to explicitly mine the conditional dependences between these distinct sparse features. Our graphical models are synthesized using simple tree structures which can be discriminatively learnt (even with limited training samples) for classification. Experiments on benchmark HSI data sets reveal significant improvements over existing approaches in classification rates as well as robustness to choice of training.**

***Index Terms*—Classification, hyperspectral imagery, joint sparsity model, probabilistic graphical models, sparse representation, spatial correlation.**

I. Introduction

HYPERSPECTRAL imaging sensors acquire digital images in hundreds of continuous narrow spectral bands spanning the visible-to-infrared spectrum [1]. A pixel in hyperspectral images (HSIs) is typically a high-dimensional vector of intensities as a function of wavelength. The high spectral resolution of the HSI pixels facilitates superior discrimination of object types.

In HSI classification, the class label of each pixel is determined given a representative training set from each class. The support vector machine (SVM) [2], which solves binary classification problems by finding the optimal separating hyperplane between the two classes, has proved to be a powerful classifier for HSI classification tasks [3]. Variants such as SVM with composite kernels, which incorporates spatial information directly in the kernels [4], have led to improved performance.

Recent work has highlighted the relevance of incorporating contextual information during HSI classification to improve performance [4]–[7], particularly because HSI pixels in a local neighborhood generally correspond to the same material and have similar spectral characteristics. Many approaches have exploited this aspect, for example, by including postprocessing of individually labeled samples [5], [6] and Markov random fields in Bayesian approaches [7]. The composite kernel approach [4] combines the spectral and spatial information from each HSI pixel via kernel composition.

An important recent advance exploits sparsity for HSI classification [8], using the observation that spectral signatures of the same material lie in a subspace of reduced dimensionality compared to the number of spectral bands. An unknown pixel is then expressed as a sparse linear combination of a few training samples from a given dictionary, and the underlying sparse representation vector encodes the class information. Furthermore, to exploit spatial correlation, a joint sparsity model is employed in [8], wherein neighboring pixels are assumed to be represented by linear combinations of a few *common* training samples to enforce smoothness across these pixels.

The technique in [8] performs classification by using (spectral) reconstruction error computed over the pixel neighborhood. Recent work [9] in model-based compressed sensing has shown the benefits of using probabilistic graphical models as priors on sparse coefficients for signal (e.g., image) reconstruction problems. Inspired by this, we propose to use probabilistic graphical models to enforce a *class-specific* structure on sparse coefficients, wherein our designed graphs represent class conditional densities. We claim that the distinct sparse representations (corresponding to each pixel in a spatial neighborhood) resulting from the joint sparsity model [8] offer *complementary yet correlated* information for classification. Our proposed framework then exploits these class conditional correlations into building a powerful classifier. Specifically, a pair of discriminative tree graphs [10] is first learnt for each distinct set of features, i.e., the sparse representation vectors of each pixel in the local spatial neighborhood of a central pixel. These initially disjoint graphs are then thickened (by introducing new edges) into a richer graphical structure via boosting [10]–[12]. The training phase of our graphical model learning uses sparse coefficients from all HSI classes, and therefore, we learn a *discriminative* graph-based classifier that captures interclass information which is ignored by the *reconstruction* residual in [8]. Evaluation on benchmark HSI data sets reveals that exploiting the structure on sparse coefficients

Manuscript received December 21, 2011; revised April 27, 2012 and July 7, 2012; accepted July 24, 2012. Date of publication September 10, 2012; date of current version November 24, 2012. This work was supported in part by the National Science Foundation under Grants CCF-1117545 and CCF-0728893, by the Army Research Office under Grants 58110-MA-II and 60219-MA, and by the Office of Naval Research under Grant N102-183-0208.

U. Srinivas and V. Monga are with the Department of Electrical Engineering, Pennsylvania State University, University Park, PA 16802 USA (e-mail: uxs113@psu.edu; vmonga@engr.psu.edu).

Y. Chen and T. D. Tran are with the Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: ychen98@jhu.edu; trac@jhu.edu).

N. M. Nasrabadi is with the U.S. Army Research Laboratory, Adelphi, MD 20783-1197 USA (e-mail: nnasraba@arl.army.mil).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2012.2211858

via class conditional graphs offers significant improvements in classification rates. Crucially, our technique exhibits a more graceful degradation with a decrease in the number of training HSI pixels, over state-of-the-art alternatives.

II. Background

A. Sparsity Model for Hyperspectral Classification

The HSI sparsity model is an extension of the sparse-representation-based framework first introduced for face recognition [13]. This model relies on the key observation that the spectral signatures of pixels approximately lie in a low-dimensional subspace spanned by representative training pixels from the same class. Consequently, for a test pixel whose class identity is unknown, there exists a sparse representation in terms of training samples from all classes. Let $\boldsymbol{y} \in \mathbb{R}^B$ be a pixel with B indicating the number of spectral bands and $\boldsymbol{D}_m \in \mathbb{R}^{B \times N_m}$, $m = 1, 2, \ldots, M$, be the subdictionary whose columns are the N_m training samples from the mth class. The HSI pixel $\boldsymbol{y}$ can then be written as

$$\boldsymbol{y} = \boldsymbol{D}_1\boldsymbol{\alpha}_1 + \cdots + \boldsymbol{D}_M\boldsymbol{\alpha}_M = \underbrace{[\boldsymbol{D}_1 \ \cdots \ \boldsymbol{D}_M]}_{\boldsymbol{D}} \underbrace{\begin{bmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_M \end{bmatrix}}_{\boldsymbol{\alpha}} = \boldsymbol{D}\boldsymbol{\alpha} \tag{1}$$

where $\boldsymbol{D} \in \mathbb{R}^{B \times N}$ with $N = \sum_{m=1}^{M} N_m$ is a structured dictionary consisting of training samples (referred to as atoms) from all classes and $\boldsymbol{\alpha} \in \mathbb{R}^N$ is a sparse vector. Given the overcomplete dictionary $\boldsymbol{D}$, the sparse coefficient vector $\boldsymbol{\alpha}$ is obtained by solving the following optimization problem:

$$\hat{\boldsymbol{\alpha}} = \arg\min \|\boldsymbol{\alpha}\|_0 \quad \text{subject to} \quad \|\boldsymbol{y} - \boldsymbol{D}\boldsymbol{\alpha}\|_2 \leq \varepsilon \tag{2}$$

where ε is a suitably chosen reconstruction error tolerance. The sparse vector $\hat{\boldsymbol{\alpha}}$ can be recovered efficiently using many norm minimization techniques, including greedy algorithms or ℓ_1-norm relaxation [14]. The class label of $\boldsymbol{y}$ is finally determined by the minimal residual between $\boldsymbol{y}$ and its approximation from each class subdictionary

$$\text{Class}(\boldsymbol{y}) = \arg \min_{m=1,\ldots,M} \|\boldsymbol{y} - \boldsymbol{D}_m\hat{\boldsymbol{\alpha}}_m\|_2 \tag{3}$$

where $\hat{\boldsymbol{\alpha}}_m$ is the collection of coefficients in $\hat{\boldsymbol{\alpha}}$ corresponding to the mth-class subdictionary.

B. Joint Sparsity Model

HSIs are usually smooth in the sense that pixels within a small neighborhood usually represent the same material, and thus, their spectral characteristics are highly correlated. In order to incorporate this spatial correlation information, the joint sparsity model [15] is employed for HSI classification in [8] by assuming that the sparse vectors associated with pixels in a local spatial neighborhood share a common sparsity pattern. Specifically, let $\{\boldsymbol{y}_t\}_{t=1,\ldots,T}$ be T pixels in a spatial neighborhood centered at $\boldsymbol{y}_1$. These neighboring pixels can be expressed as

$$\begin{aligned} \boldsymbol{Y} &= [\boldsymbol{y}_1 \quad \boldsymbol{y}_2 \quad \cdots \quad \boldsymbol{y}_T] = [\boldsymbol{D}\boldsymbol{\alpha}_1 \quad \boldsymbol{D}\boldsymbol{\alpha}_2 \quad \cdots \quad \boldsymbol{D}\boldsymbol{\alpha}_T] \\ &= \boldsymbol{D}\underbrace{[\boldsymbol{\alpha}_1 \quad \boldsymbol{\alpha}_2 \quad \cdots \quad \boldsymbol{\alpha}_T]}_{\boldsymbol{S}} = \boldsymbol{D}\boldsymbol{S}. \end{aligned} \tag{4}$$

The sparse vectors $\{\boldsymbol{\alpha}_t\}_{t=1,\ldots,T}$ share the same support, i.e., they are linear combinations of the same collection of atoms from $\boldsymbol{D}$ but with possibly different weights assigned to each atom. As a result, $\boldsymbol{S}$ is a sparse matrix with only a few nonzero rows. This row-sparse matrix $\boldsymbol{S}$ can be recovered by solving the following constrained optimization problem:

$$\hat{\boldsymbol{S}} = \arg\min \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{S}\|_F \quad \text{subject to} \quad \|\boldsymbol{S}\|_{\text{row},0} \leq K_0 \tag{5}$$

where $\|\boldsymbol{S}\|_{\text{row},0}$ denotes the number of nonzero rows of $\boldsymbol{S}$ and $\|\cdot\|_F$ is the Frobenius norm. The problem in (5) can be approximately solved by the greedy simultaneous orthogonal matching pursuit (SOMP) algorithm [15]. The identity of $\boldsymbol{y}_1$ is then determined by the minimal total residual

$$\text{Class}(\boldsymbol{y}_1) = \arg \min_{m=1,\ldots,M} \|\boldsymbol{Y} - \boldsymbol{D}_m\hat{\boldsymbol{S}}_m\|_F \tag{6}$$

where $\hat{\boldsymbol{S}}_m$ contains the rows of $\hat{\boldsymbol{S}}$ associated with $\boldsymbol{D}_m$.

C. Probabilistic Graphical Models

A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a collection of nodes $\mathcal{V} = \{v_1, \ldots, v_r\}$ and a set of (undirected) edges $\mathcal{E} \subset \binom{\mathcal{V}}{2}$. A probabilistic graphical model describes the joint distribution of a random vector with each node representing one (or a group of) random variable(s) whose conditional dependences are indicated by the presence of connecting edges. The graph structure leads to a factorization of the joint probability distribution of the random vector in terms of marginal and pairwise statistics. The Hughes phenomenon [16] highlights the difficulty of learning models for high-dimensional data with limited number of training samples. The use of probabilistic graphs reduces sensitivity to choice of training, particularly in the low-training regime [17, Ch. 8], [18].

Graphical models can be learnt either generatively or discriminatively. In the generative setting, a single graph which approximates a given distribution is learnt by minimizing the approximation error. The seminal contribution in this area is due to Chow and Liu [19], who obtained the optimal tree approximation $\hat{p}$ of a multivariate distribution p by minimizing the Kullback–Leibler (KL) distance $D(p\|\hat{p}) = E_p[\log(p/\hat{p})]$ using first- and second-order statistics, via a maximum-weight spanning tree (MWST) problem. In discriminative learning, a pair of graphs is jointly learnt by minimizing the classification error. Recently, a discriminative learning framework has been proposed [10] by maximizing the tree-approximate J-divergence (a symmetric extension of the KL distance)

$$\hat{J}(\hat{p}, \hat{q}; p, q) = \int (p(x) - q(x)) \log \left[\frac{\hat{p}(x)}{\hat{q}(x)}\right] dx. \tag{7}$$

Based on the observation that maximizing the J-divergence minimizes the upper bound on the probability of classification error, the discriminative learning problem then becomes

$$(\hat{p}, \hat{q}) = \arg \max_{\hat{p},\hat{q} \text{ are trees}} \hat{J}(\hat{p}, \hat{q}; \tilde{p}, \tilde{q}) \tag{8}$$

where $\tilde{p}$ and $\tilde{q}$ are the empirical estimates of p and q, respectively. The problem in (8) is shown to decouple into two MWST

problems [10]

$$\hat{p} = \arg \min_{\hat{p} \text{ is a tree}} D(\tilde{p}\|\hat{p}) - D(\tilde{q}\|\hat{p})$$
$$\hat{q} = \arg \min_{\hat{q} \text{ is a tree}} D(\tilde{q}\|\hat{q}) - D(\tilde{p}\|\hat{q}). \quad (9)$$

The optimal choice of $\hat{p}$ ($\hat{q}$) simultaneously minimizes its distance to $\tilde{p}$ ($\tilde{q}$) and maximizes its distance from $\tilde{q}$ ($\tilde{p}$).

III. Exploiting Joint Sparsity via Probabilistic Graphical Models

Learning Thicker Graphical Models

The solution to (9) learns *tree-structured* graphs $\hat{p}$ and $\hat{q}$, which, albeit learned optimally, can model only a small family of distributions due to their simple edge structure. However, optimally learning complex graphical models is, in general, NP-hard [20]. This problem is practically addressed by boosting simpler graphs [10]–[12] into richer structures. Recently, we have proposed a feature fusion framework [21] for image classification where the "initial graphical structure" for boosting is chosen as a forest of disjoint tree graphs. Thickening this forest with new edges is hence tantamount to discovering new conditional correlations between distinct feature sets.

Algorithm 1 LSGM (steps 1–4 offline)

1: **Feature extraction (training)**: Compute sparse representations $\boldsymbol{\alpha}_l$, $l = 1, \ldots, T$, for neighboring pixels of the training data.
2: **Initial disjoint graphs**:
Discriminatively learn T pairs of N-node tree graphs $\mathcal{G}_l^p$ and $\mathcal{G}_l^q$ on $\{\boldsymbol{\alpha}_l\}$, for $l = 1, \ldots, T$, obtained from training data.
3: Separately concatenate nodes corresponding to the two classes, to generate initial graphs.
4: **Boosting on disjoint graphs**: Iteratively thicken initial disjoint graphs via boosting to obtain final graphs $\mathcal{G}^p$ and $\mathcal{G}^q$.

{**Online process**}
5: **Feature extraction (test)**: Obtain sparse representations $\boldsymbol{\alpha}_l$, $l = 1, \ldots, T$, in $\mathbb{R}^N$ from test image.
6: **Inference**: Classify based on the output of the resulting classifier using (10).

We leverage the framework in [21] for HSI classification. Here, the generation of distinct and complementary feature sets is achieved by solving the joint sparse recovery problem in (5) (with distinct feature sets being the sparse coefficient vectors/columns of $\boldsymbol{S}$). This is shown in Fig. 1, and the formal description of our proposed local sparsity graphical model (LSGM) algorithm is provided in Algorithm 1.

Note that the proposed LSGM algorithm consists of an *offline* training stage (steps 1–4) and an *online* classification stage (steps 5 and 6). The local sparsity in the name is indicative of the joint sparsity model that is used to obtain the local sparse features. The discriminative graphs are learnt in the training stage. The process described here is for binary classification. The approach extends to multiclass problems by learning graphs in a one-against-all manner. For an M-class classification problem, we learn M pairs of discriminative graphs that represent the class conditional probability density functions $f(\boldsymbol{\alpha}|C_m)$ and $f(\boldsymbol{\alpha}|\tilde{C}_m)$ for $m = 1, 2, \ldots, M$, where C_m denotes the mth class and $\tilde{C}_m$ denotes the complement of C_m (i.e., $\tilde{C}_m = \bigcup_{k=1,\ldots,M,k\neq m} C_k$).

We first obtain the feature vectors (i.e., sparse vectors with respect to a given training dictionary $\boldsymbol{D}$) of training samples and their neighboring pixels by solving the joint sparse recovery problem in (5). Let T be the size of the neighborhood. The extraction of sparse features may be viewed as a transformation $\mathcal{T}_l : \mathbb{R}^B \mapsto \mathbb{R}^N$, and there are T such distinct transformations $\mathcal{T}_l$, $l = 1, 2, \ldots, T$. For every pixel $\boldsymbol{y} \in \mathbb{R}^B$, T different features $\boldsymbol{\alpha}_l \in \mathbb{R}^N, l = 1, 2, \ldots, T$, are obtained, as shown in Fig. 1 for a 3×3 neighborhood with $T = 9$ (only three features are displayed). For each type of feature, training features for class C_m correspond to pixels in a neighborhood of training samples known to belong to class C_m. Features for $\tilde{C}_m$ are the sparse vectors associated with neighbors of representative training.

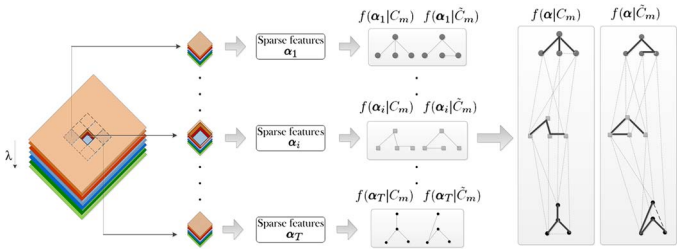


Fig. 1. HSI classification using discriminative graphical models on sparse feature representations obtained from local pixel neighborhoods.

For each of the T transformations $\mathcal{T}_l$, a pair of N-node discriminative tree graphs $\mathcal{G}_l^p$ and $\mathcal{G}_l^q$, which approximate the class distributions $f(\boldsymbol{\alpha}_l|C_m)$ and $f(\boldsymbol{\alpha}_l|\tilde{C}_m)$, respectively, is simultaneously learnt by solving the decoupled MWST problems in (9). The initial disjoint graphs with TN nodes representing the class distribution corresponding to C_m and $\tilde{C}_m$ are then generated by separately concatenating the nodes of $\mathcal{G}_l^p$, $l = 1, \ldots, T$, and $\mathcal{G}_l^q$, $l = 1, \ldots, T$, respectively. These graphs with sparse edge structure are then iteratively thickened via boosting [21]. Different pairs of discriminative graphs over the same sets of nodes with different weights are learnt in different iterations, and the newly learnt edges are used to augment the graphs. The final "thickened" graphs $\mathcal{G}^p$ and $\mathcal{G}^q$ are shown in Fig. 1 (right side).

The process described earlier (steps 1–4 in Algorithm 1) is performed offline, and M pairs of discriminative graphs are learnt for the M binary classification problems in a one-against-all manner. The classification of a new test sample is then performed online. Features $\boldsymbol{\alpha}$ are extracted from the test sample $\boldsymbol{y}$ by solving the sparse recovery problem in (5) for the T pixels in the neighborhood centered at $\boldsymbol{y}$. Let $\hat{f}(\boldsymbol{\alpha}|C_m)$ and $\hat{f}(\boldsymbol{\alpha}|\tilde{C}_m)$ denote the final graphs learnt for C_m and $\tilde{C}_m$, respectively. The class label of $\boldsymbol{y}$ is determined as follows:

$$\text{Class}(\boldsymbol{y}) = \arg \max_{m \in \{1,\ldots,M\}} \log \left(\frac{\hat{f}(\boldsymbol{\alpha}|C_m)}{\hat{f}(\boldsymbol{\alpha}|\tilde{C}_m)} \right). \quad (10)$$

IV. Experiments and Results

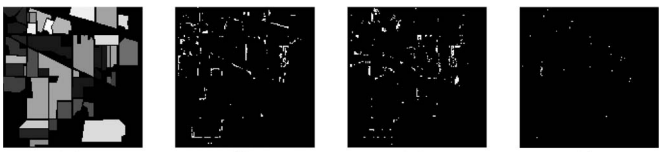
We compare our proposed LSGM approach with three competitive methods: 1) spectral-feature-based SVM classifier [3], [22]; 2) composite kernel SVM (SVM-CK) [4]; and 3) joint sparsity model (SOMP) [8]. In SVM-CK, two types of kernels are used: a spectral kernel K_ω for the spectral (pixel) features (in $\mathbb{R}^{200}$) and a spatial kernel K_s for the spatial features (in $\mathbb{R}^{400}$) which are formed by the mean and standard deviation of pixels in a neighborhood per spectral channel. A polynomial kernel (order $d = 7$) is used for the spectral features, while the radial basis function (RBF) kernel is used

TABLE I
CLASSIFICATION RATES FOR THE AVIRIS INDIAN PINES TEST SET. LSGMz-SCORE$= -2.13$

Class type	Training	Test	SVM	SVM-CK	SOMP	LSGM
Alfalfa	6	48	83.33	95.83	87.50	89.58
Corn-notill	144	1290	88.06	96.82	94.80	95.50
Corn-min	84	750	72.40	91.20	94.53	94.80
Corn	24	210	60.48	87.62	93.33	94.76
Grass/pasture	50	447	92.39	93.74	89.71	90.82
Grass/trees	75	672	96.72	97.62	98.51	99.55
Pasture-mowed	3	23	47.82	73.91	91.30	91.30
Hay-windrowed	49	440	98.41	98.86	99.32	99.55
Oats	2	18	50.00	55.56	0	44.44
Soybeans-notill	97	871	72.91	94.26	89.44	90.93
Soybeans-min	247	2221	85.14	94.73	97.03	97.39
Soybeans-clean	62	552	86.23	93.84	88.94	92.39
Wheat	22	190	99.47	99.47	100	100
Woods	130	1164	93.73	99.05	99.57	99.65
Building-trees	38	342	63.45	88.01	98.83	99.71
Stone-steel	10	85	87.05	100	97.65	98.82
Overall	1043	9323	85.11	95.15	95.31	96.18

TABLE II
CLASSIFICATION RATES FOR THE UNIVERSITY OF PAVIA TEST SET. LSGMz-SCORE$= -2.01$

Class type	Training	Test	SVM	SVM-CK	SOMP	LSGM
Asphalt	548	6304	84.01	80.20	59.49	66.55
Meadows	540	18146	67.50	84.99	78.31	86.10
Gravel	392	1815	67.49	82.37	84.13	86.72
Trees	524	2912	97.32	96.33	96.30	96.94
Metal sheets	265	1113	99.28	99.82	87.78	98.83
Bare soil	532	4572	92.65	93.35	77.45	94.62
Bitumen	375	981	89.70	90.21	98.67	99.18
Bricks	514	3364	92.24	92.95	89.00	94.44
Shadows	231	795	96.73	95.85	91.70	96.10
Overall	3921	40002	79.24	87.33	78.75	86.38



(a) (b) (c) (d)

Fig. 2. Difference maps for the AVIRIS Indian Pines data set, for the ground truth map in (a). (b) SVM-CK [4]. (c) SOMP [8]. (d) Proposed LSGM approach.

for the spatial features. The σ parameter for the RBF kernel and the SVM regularization parameter C are selected by cross-validation. The weighted summation kernel $K = \mu K_s + (1-\mu)K_\omega$ effectively captures spectral and contextual spatial information, with the optimal choice $\mu = 0.4$ determined by cross-validation. A 5×5 window is used for the neighborhood kernels. The parameters for SOMP are chosen as described in [8]. The proposed LSGM approach uses a local window of dimension 3×3. For fairness of comparison, results for SOMP are also presented for the same window dimension.

We perform experiments using three distinct HSI data sets. Note that two flavors of the results are reported: 1) Tables I–III show the classification rates for carefully selected or good training samples which amount to about 10% of available data (typical of training choices in [4] and [8]), and 2) Fig. 3(a)–(f) shows the performance plotted as a function of training set size and also the results averaged from multiple (ten) random training runs. Fig. 3(d)–(f) characterizes the distribution of the classification rates (modeled as a random variable whose value emerges as an outcome of a given run and fit to a Gaussian distribution). Furthermore, we establish the statistical significance of our results by computing LSGM z-scores for each data set.

TABLE III
CLASSIFICATION RATES FOR THE CENTER OF PAVIA TEST SET. LSGMz-SCORE$= -2.17$

Class type	Training	Test	SVM	SVM-CK	SOMP	LSGM
Water	745	64533	99.19	97.61	99.38	99.44
Trees	785	5722	77.74	92.99	91.98	92.99
Meadow	797	2094	86.72	97.37	95.89	96.99
Brick	485	1667	40.37	79.60	86.44	87.28
Soil	820	5729	97.52	98.65	96.75	97.64
Asphalt	678	6847	94.77	94.37	93.79	94.54
Bitumen	808	6479	74.37	97.53	95.06	96.99
Tile	223	2899	98.94	99.86	99.83	99.90
Shadow	195	1970	100	99.89	98.48	99.34
Overall	5536	97940	94.63	96.97	97.82	98.20

A. AVIRIS Data Set: Indian Pines

The first HSI in our experiments is the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) Indian Pines image [23]. The AVIRIS sensor generates 220 bands across the spectral range from 0.2 to 2.4 μm, of which only 200 bands are considered by removing 20 water absorption bands [22]. This image has a spatial resolution of 20 m per pixel and a spatial dimension of 145×145. For well-chosen training samples, the difference maps obtained using the different approaches are shown in Fig. 2(b)–(d), and the classification rates for each class, as well as the overall accuracy, are shown in Table I. The improvement over SOMP indicates the benefits of using a discriminative classifier instead of reconstruction residuals for class assignment while still retaining the advantages of exploiting spatiospectral information.

Fig. 3(a) shows a comparison of algorithm performances as a function of training set size. Our LSGM approach outperforms the competing approaches, and the difference is particularly significant in the low-training regime. As expected, the overall classification accuracy decreases when the number of training samples is reduced. With that said, LSGM offers a more graceful degradation in comparison to other approaches. In Fig. 3(d), the average classification rate is the highest for LSGM, which is consistent with the results in Fig. 3(a). Furthermore, the variance is the lowest for LSGM, underlining its improved robustness against particular choice of training samples.

B. ROSIS Urban Data Over Pavia, Italy

The next two HSIs, University of Pavia and Center of Pavia, are urban images acquired by the Reflective Optics System Imaging Spectrometer (ROSIS). The ROSIS sensor generates 115 spectral bands ranging from 0.43 to 0.86 m and has a spatial resolution of 1.3 m per pixel. The University of Pavia image consists of 610×340 pixels, each having 103 bands with the 12 noisiest bands removed. The Center of Pavia image consists of 1096×492 pixels, each having 102 spectral bands after 13 noisy bands are removed. For these two images, we repeat the experimental scenarios tested in Section IV-A.

The classification rates for the two ROSIS images are provided in Tables II and III, respectively, for the scenario of well-chosen training samples. In Table II, the SVM-CK technique performs marginally better than LSGM in the sense of overall classification accuracy. However, for most individual classes, LSGM does better, particularly in cases where the training sample size is smaller. In Table III, LSGM performs better than SOMP as well as SVM-CK. From Fig. 3(b) and (c), we observe that LSGM improves upon the performance of SOMP

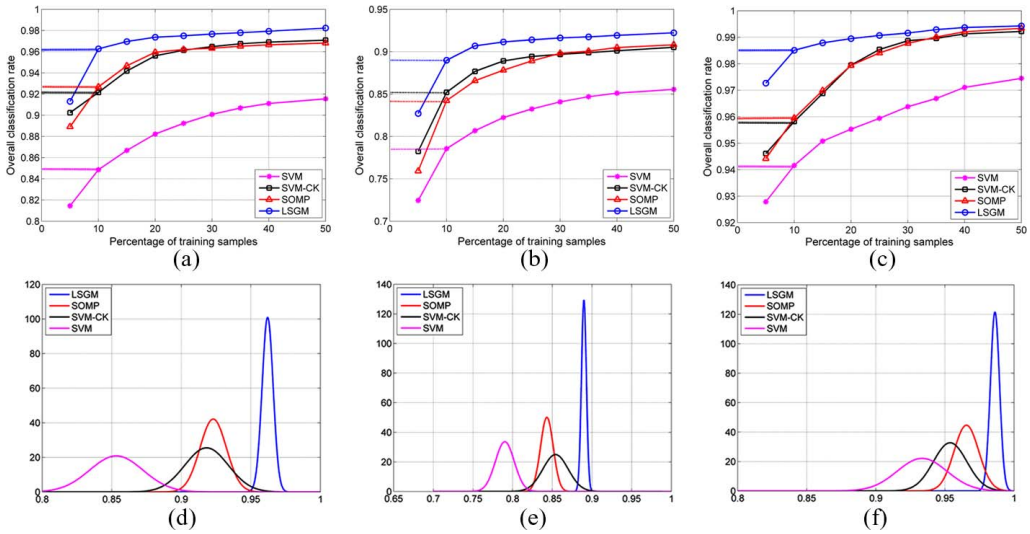


Fig. 3. Performance of different approaches as a function of the number of training samples provided. (a) AVIRIS image. (b) University of Pavia image. (c) Center of Pavia image. (d)–(f) Density function of the classification rates obtained for ten different random realizations of training.

and SVM-CK by about 4%, while the improvements over the baseline SVM classifier are even more pronounced.

The z-score for LSGM on the AVIRIS image is -2.13, which indicates that, with a high probability $(= 0.983)$, any random selection of training samples will give results similar to the values in Table I. For the University of Pavia and Center of Pavia images, the z-scores are -2.01 and -2.17, respectively. The negative sign merely indicates that the experimental value is lesser than the most likely value (Gaussian mean).

V. Conclusion

Linear reconstruction models that impose sparsity are gaining increasing popularity in HSI classification. A spatiospectral notion of sparsity is exploited in the current work by posing a structure on sparse coefficients via discriminative graphical models. Results show marked improvement over powerful state-of-the-art classifiers, particularly in the form of robustness to choice and number of training hyperspectral profiles.

Acknowledgment

The authors would like to thank the University of Pavia and the HySenS project for kindly providing the Reflective Optics System Imaging Spectrometer images of University of Pavia and Center of Pavia.

References

[1] M. Borengasser, W. S. Hungate, and R. Watkins, *Hyperspectral Remote Sensing—Principles and Applications*. Boca Raton, FL: CRC Press, 2008.

[2] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Berlin, Germany: Springer-Verlag, 1995.

[3] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

[4] G. Camps-Valls, L. Gomez-Chova, J. Muñoz-Marí, J. Vila-Francés, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 1, pp. 93–97, Jan. 2006.

[5] F. Bovolo, L. Bruzzone, and M. Marconcini, "A novel context-sensitive SVM for classification of remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Denver, Jul. 2006, pp. 2498–2501.

[6] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot, "Spectral–spatial classification of hyperspectral imagery based on partitional clustering techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 8, pp. 2973–2987, Aug. 2009.

[7] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Spectral–spatial hyperspectral image segmentation using subspace multinomial logistic regression and Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 3, pp. 809–823, Mar. 2012.

[8] Y. Chen, N. Nasrabadi, and T. D. Tran, "Hyperspectral image classification using dictionary-based sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3973–3985, Oct. 2011.

[9] V. Cevher, P. Indyk, L. Carin, and R. Baraniuk, "Sparse signal acquisition and recovery with graphical models," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 92–103, Nov. 2010.

[10] V. Y. F. Tan, S. Sanghavi, J. W. Fisher, and A. S. Willsky, "Learning graphical models for hypothesis testing and classification," *IEEE Trans. Signal Process.*, vol. 58, no. 11, pp. 5481–5495, Nov. 2010.

[11] Y. Freund and R. E. Schapire, "A short introduction to boosting," *J. Jpn. Soc. Artif. Intell.*, vol. 14, no. 5, pp. 771–780, Sep. 1999.

[12] T. Downs and A. Tang, "Boosting the tree augmented naive Bayes classifier," in *Proc. Intell. Data Eng. Automated Learn., Lecture Notes in Computer Science*, 2004, vol. 3177, pp. 708–713.

[13] J. Wright, A. Y. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[14] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Rev.*, vol. 51, no. 1, pp. 34–81, Feb. 2009.

[15] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Process.*, vol. 86, no. 3, pp. 572–588, Mar. 2006.

[16] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 1, pp. 55–63, Jan. 1968.

[17] C. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006.

[18] M. J. Wainwright and M. I. Jordan, *Graphical Models, Exponential Families, and Variational Inference*. Hanover, MA: Now Publ. Inc., 2008.

[19] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 3, pp. 462–467, May 1968.

[20] G. F. Cooper, "The computational complexity of probabilistic inference using Bayesian belief networks," *Artif. Intell.*, vol. 42, no. 2/3, pp. 393–405, Mar. 1990.

[21] U. Srinivas, V. Monga, and R. G. Raj, "Automatic target recognition using discriminative graphical models," in *Proc. IEEE Int. Conf. Image Process.*, 2011, pp. 33–36.

[22] J. A. Gualtieri and R. F. Cromp, "Support vector machines for hyperspectral remote sensing classification," in *Proc. SPIE*, Jan. 1998, vol. 3584, pp. 221–232.

[23] AVIRIS NW Indiana's Indian Pines 1992 Data Set. [Online]. Available: http://cobweb.ecn.purdue.edu/biehl/MultiSpec/documentation.html